Helsinki University of Technology, Department of Chemical Technology
Espoo 2001 TKK-BE-3

Technical Biochemistry Report 1/2001

COMPUTATIONAL ANALYSIS OF PROTEIN FUNCTION - LIPASES AND A-KINASE ANCHORING PROTEINS

Sanna Herrgård

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Chemical Technology for public examination and debate in Auditorium KE 2 (Komppa Auditorium) at Helsinki University of Technology (Espoo, Finland) on the 16th of March, 2001, at 12 noon.

Helsinki University of Technology
Department of Chemical Technology
Laboratory of Bioprocess Engineering

Teknillinen korkeakoulu
Kemian tekniikan osasto
Bioprosessitekniikan laboratorio

Distribution:
Helsinki University of Technology
Laboratory of Bioprocess Engineering
P.O. Box 6100
FIN-02015 HUT
Tel. +358-9-4512541
Fax. +358-9-462373
Email: sannaherrgard@geneformatics.com

ISBN 951-22-5358-5
ISSN 0359-6621

Otamedia Oy
Espoo 2001

Herrgård, Sanna. Computational analysis of protein function - lipases and A-kinase anchoring proteins. Espoo 2001, Helsinki University of Technology.

Keywords: A-kinase anchoring protein, domain architecture, domain motion, electrostatics, normal mode analysis, *Rhizomucor miehei* lipase.

Abstract

In this thesis, normal mode analysis, electrostatic calculations, and sequence-based bioinformatics methods were applied to investigate protein sequence-structure-function relationship. Two different protein families were studied: a family of fungal lipases (the *Rhizomucor miehei* lipase or RmL family), and a family of adapter proteins called A-kinase anchoring proteins (AKAPs). The RmL family is well conserved, and three-dimensional structures for several members of this family are known. The AKAP family is a functionally, but not structurally, related protein family that consists of multi-domain proteins whose domain architectures are not well understood.

In the first part of this thesis, lipase structure-function relationship was investigated using two complementary methods: normal mode analysis and electrostatic calculations. Normal mode analysis was applied to characterize the collective motions in RmL; a global breathing motion as well as local loop motions were shown to be associated with the conformational change. The electrostatic calculations yield a detailed description of the role of electrostatic interactions in the RmL structure, function, and stability. Two main results were obtained in this study. First, the key residues that affect the lid stability were identified. Second, a network of electrostatic interactions was discovered. This network is an important feature of the lipase structure, since it connects the active site to the mobile lid region. The network was observed to be conserved in the RmL family and their homologues.

In the second part of the thesis, the sequence-function relationship of an AKAP from *Caenorhabditis elegans* ($\mathrm{AKAP_{CE}}$) was studied using sequence-based bioinformatics methods. The domain architecture of $\mathrm{AKAP_{CE}}$ reveals that it shares two domains with SARA (Smad anchor for receptor activation), which is an adapter protein involved in the TGFβ signaling pathway. One of the domains, FYVE, is well characterized, whereas the other one, a TGFβ receptor binding domain, has not been characterized earlier. The existence of these two domains in $\mathrm{AKAP_{CE}}$ leads us to propose a novel AKAP function as a TGFβ receptor binding protein.

Preface

This thesis is a compilation of work carried out at Helsinki University of Technology, University of York (England), University of Illinois at Urbana-Champaign (USA), and University of California, San Diego (USA). I have been a student in the Finnish Graduate School on Applied Bioscience: Bio-engineering, Food & Nutrition, Environment. The work was carried out during years 1995-2000.

Literally speaking, doing this thesis has been quite a journey. When I started graduate school, I never guessed that I would end up in Southern California via England and the Midwest. I consider myself very lucky in that I have had the opportunity to work with so many wonderful people in Finland, England, and in the United States.

I would like to express my gratitude to Professor emeritus Pekka Linko, Dr. Yu-Yen Linko and Dr. Susan Linko from the Laboratory of Bioprocess Engineering as well as Dr. Tuija Raaska and Dr. Leif Laaksonen from the Center for Scientific Computing for their enthusiasm and encouragement in the initial stages of this work. I want to thank especially Yu-Yen for suggesting that investigating lipase structure could be helpful in understanding their function. Professor Matti Leisola is warmly acknowledged for being my supervisor, and giving me the freedom and opportunity to pursue my scientific interests. I would also like to thank my colleagues at HUT for their help, friendship and support. I would especially like to acknowledge Dr. Xiao Yan Wu for her help and collaboration.

I spent the academic year 1996-1997 and part of 1997-1998 in the Protein Structural Biology Laboratory in York. I am very grateful for Professor Rod Hubbard, Dr. Leo Caves and Dr. Chandra Verma for teaching me a lot about protein structures and computational structural biology. My warm thanks are due to Leo for teaching me how to write scientific papers. I won't forget the frustration and the sweat involved in writing the first paper of this thesis, but I think that it did teach me quite a bit about scientific writing. As Professor Russell Doolittle wrote in his recent paper in Bioinformatics [1] *"Science is not all thrill and satisfaction. There is always disappointment and frustration, and occasionally some outright pain"*. However, I feel that there is often something good that comes out of this frustration. I want to thank Chandra for his patience in answering the many, many questions I had every day, for being my friend, and for reminding me regularly that life is not so serious, afterall. Professor Guy Dodson and Mrs. Eleanor Dodson are warmly thanked for their

efforts in making the group so special.

I moved from Finland to the United States in Fall 1998 to continue my research in Professor Shankar Subramaniam's group. I wish to express my gratitude to Shankar for inspiring research ideas, constructive feedback on my work, and reminding me that even if you use theoretical methods to investigate biological systems, you need to have a biological question in your mind that you are trying to answer. My warmest thanks are due to Professor Cynthia Gibas for her help with the electrostatic methods, and for Per Jambeck for introducing me to the world of bioinformatics. Professor Susan Taylor from the Department of Chemistry and Biochemistry at UCSD is warmly acknowledged for a fruitful collaboration. I would also like to thank Signal Pharmaceuticals for their collaboration.

The two pre-examiners of this thesis, Professor Mark Johnson from Åbo Akademi University and Professor Mauno Vihinen from University of Tampere, are warmly thanked for their constructive feedback and good suggestions.

I would like to acknowledge the following institutions and foundations for financial support: the ABS graduate school, the Academy of Finland, the Alfred Kordelin Foundation, the European Union, the Food Research Foundation, the Foundation of Technology, the Helsinki University of Technology Foundation for Financial Aid, and the Leo and Regina Wainstein Foundation.

Last, and most importantly, I thank my husband Markus for his love and support (which sometimes meant cooking and cleaning for him when I was writing this thesis), and for sharing so many moments of joy in my life. I would also like to thank my parents Manu and Riitta, my brother Panu and my sister Anna, and my many friends for their support and understanding.

La Jolla, February 10, 2001

Sanna Herrgård

List of publications

This thesis is based on the following publications which are referred as I to IV in the text.

I S. Jääskeläinen, C.S. Verma, R.E. Hubbard, P. Linko, and L.S.D. Caves. Conformational change in the activation of lipase: an analysis in terms of low-frequency normal modes. *Protein Sci.*, 7:1359-1367, 1998.

II S. Jääskeläinen, C.S. Verma, R.E. Hubbard, and L.S.D. Caves. Identifying key electrostatic interactions in *Rhizomucor miehei* lipase: the influence of solvent dielectric. *Theor. Chem. Acc.*, 101:175-179, 1999.

III S. Herrgård, C.J. Gibas, and S. Subramaniam. Role of an electrostatic network of residues in the enzymatic action of the *Rhizomucor miehei* lipase family. *Biochemistry*, 39:2921-2930, 2000.

IV S. Herrgård, P. Jambeck, S.S. Taylor, and S. Subramaniam. Domain architecture of a *Caenorhabditis elegans* AKAP suggests a novel AKAP function. *FEBS Lett.*, 486:107-111, 2000.

Abbreviations

AKAP	A-kinase anchoring protein
$AKAP_{CE}$	A-kinase anchoring protein from *Caenorhabditis elegans*
CDS	coding sequence
C. elegans	*Caenorhabditis elegans*
DIMB	diagonalization in a mixed basis
FD	finite-difference
FRET	fluorescence resonance energy transfer
HMM	hidden Markov model
MD	molecular dynamics
MEME	multiple expectation-maximization for motif elicitation
NCBI	National Center for Biotechnology Information
NMA	normal mode analysis
NMR	nuclear magnetic resonance
PB	Poisson-Boltzmann
PDB	Protein Data Bank
PIR	Protein Information Resource
PKA	cAMP-dependent protein kinase
PLA2	phospholipase A_2
PRF	Protein Research Foundation
R. miehei	*Rhizomucor miehei*
RmL	*Rhizomucor miehei* lipase
RMS	root mean square
R. niveus	*Rhizopus niveus*
RnL	*Rhizopus niveus* lipase
SARA	Smad anchor for receptor activation
SCOP	Structural Classification of Proteins
spFRET	single pair fluorescence resonance energy transfer
TGFβ	transforming growth factor β
T. lanuginosus	*Thermomyces lanuginosus*
UHBD	University of Houston Brownian dynamics
Å	Ångström, 10^{-10} m

Contents

Chapter 1

Introduction

1.1 Protein sequence-structure-function relationship

The protein sequence-structure-function relationship forms the central paradigm of biology. Protein sequence folds into a structure, and protein structure determines its function. Understanding protein sequence and structure has proven to be essential for understanding protein function. The purpose of this introduction is to summarize the ways in which computational methods can be used to investigate and explain protein sequence, structure and function, to predict novel protein functions, and to formulate experimentally testable hypotheses.

Before a protein sequence is available for analysis, it has often undergone processing by multiple computational methods, especially if the protein was predicted from genomic DNA. These pre-processing steps include assembly of DNA (see e.g. [2]), prediction of the gene structure [3], and partial annotation of the functions of the proteins coded by the genes (see e.g. [4]).

Sequence-based bioinformatics methods are used to analyze protein and DNA sequences. Sequence comparison methods are based on an idea that if two sequences show a high degree of similarity, they are likely to have at least a similar fold and in some cases also a similar function. The assumption that proteins with significant sequence similarity have similar structures is based on the fact that protein structures are generally better conserved than their sequences [5]. The question whether two sequences with a high degree of sequence similarity also have a similar function can be addressed by investigating, whether catalytic amino acids and other key residues have been conserved. Computational functional annotation of protein sequences aims at determining the function of an unknown sequence by finding similar sequences, domains and sequence motifs whose functions are known [6]. When annotating functions of multi-domain proteins, function of each individual domain should be annotated, because each domain can reveal a new functionality for the protein [6, 7].

Sequence-based bioinformatics methods are useful not only for finding sequence similarities, but also for investigating protein families. Sequence alignments of protein families can be used to identify evolutionarily conserved regions in proteins [8, 9, 10, 11]. These regions are often likely to be crucial for the structure or function of the protein.

Polypeptide chains fold into three-dimensional structures. Understanding the folding mechanism of a protein sequence to its structure is one of the key challenges of computational biology, and multiple methods including lattice simulations and all-atom protein models with explicit or implicit solvent have been used in the effort of elucidating the protein folding mechanism [12, 13, 14]. Although the folding mechanism is not yet fully understood, it is possible to predict the secondary [15, 16] and tertiary structures of proteins from their sequences at a reasonable accuracy. Three main methods exist for predicting protein three-dimensional structures from their sequences: homology modeling [17, 18], threading [19, 20], and ab initio methods [21].

Once the three-dimensional structure of a protein is known, its structural neighbors can be identified by searching structure databases. Classification of protein structures and domains into families, superfamilies, folds and classes aims at dividing the protein structure universe into discrete groups [22, 23]. According to the current estimates there are between 1000 and 6000 different protein folds, of which almost 700 have been characterized [24, 22].

When the three-dimensional structure of a protein is known, computational methods can be applied to investigate its surface properties and the short- and long-range interactions that largely determine protein flexibility and its interactions with ligands. The main methods to investigate the static and dynamic properties of proteins are based on molecular mechanics and molecular dynamics in which a potential energy function is used to describe the interactions between the atoms in a protein [25]. Molecular dynamics (MD) calculations use Newton's laws of motion to determine the positions and velocities of the atoms as a function of time [26]. The disadvantage of the MD simulation methodology is that, due to the high computational cost, MD simulations are often not able to describe the collective motions in proteins that typically occur on timescales greater than nanoseconds [27, 28]. A method that can be used to investigate the collective motions in proteins is normal mode analysis, in which the low-frequency vibrations of a system are computed directly from the potential energy function [29].

Electrostatic interactions are the most important long-range interactions in proteins. Methods based on numerical solutions to the Poisson-Boltzmann equation can be used to study these interactions in proteins [30, 31]. Electrostatic calculations have been extensively used to investigate the role of electrostatic interactions in protein structure, function and stability, and to study protein-ligand interactions [30, 32, 33, 34, 35, 36, 37, 38].

As this introduction shows, a large number of computational methods can be used to investigate protein sequence, structure and function. The choice of the method depends on two factors: the available data, and the biological

problem one is trying to solve. If no structure of the protein is available, computational studies must be restricted to the sequence level, and the structure of the protein can be predicted, if needed. If the structure of the protein is known, a combination of the sequence-based and structure-based approaches can be used.

In this thesis, computational methods were used to investigate the protein sequence-structure-function relationship in two model systems: lipases and AKAPs. In the lipase part, the structure and function of the fungal *Rhizomucor miehei* lipase (RmL) family was investigated. All members of this family share a significant degree of conservation at the sequence and structural level, and structures for several members of this family are known. Available experimental mutational data was used both to guide the design of the computational studies, and to validate the findings. Normal mode analysis and electrostatic calculations were performed to characterize the conformational change, as well as to identify the key electrostatic interactions in the RmL family. Further, the role of the conserved residues and sequence motifs for the lipase structure and function was considered.

In the second part of the thesis, a family of functionally, but not structurally, conserved adapter proteins called AKAPs (A-kinase anchoring proteins) was studied. AKAPs consist of multiple domains, but their domain architecture is poorly understood. Since very little is currently known about the three-dimensional structures of AKAPs, sequence-based bioinformatics methods were used to investigate domain architectures of AKAPs, to identify novel domains, and to predict novel AKAP functions. In this thesis, an analysis of the domain architecture of an AKAP from the roundworm *Caenorhabditis elegans* is presented.

1.2 Lipases

Lipases (glycerol-ester-hydrolase, EC 3.1.1.3) are enzymes that catalyze hydrolysis of acylglycerols into fatty acids and glycerol, and, thus have an important role in lipid metabolism [39]. In vertebrates, different types of lipases control digestion, absorption, and reconstitution of fat, as well as lipoprotein metabolism [39, 40]. In plants, lipase activity has been identified for instance in seeds during germination [41]. Many microorganisms have been observed to produce intra- and extracellular lipases [42]. In this thesis, the structure and function of triacylglycerol lipases were investigated. Thus, 'lipase' refers here specifically to a triacylglycerol lipase.

Lipases are widely used in industrial applications: in surfactants, detergents, the dairy industry, pulp and the paper industry and in oleochemical products [43, 44]. The ability of lipases to catalyze esterification reactions in organic solvents is used in the synthesis of structured triglycerides (e.g. cocoa butter substitutes [45, 46]) and in the production of polymers [47]. Further, lipases can also be used to produce optically pure compounds for the pharma-

ceutical industry [48, 49].

A notable characteristic of lipases is the increase in their activity when the substrate concentration exceeds the critical micelle concentration [50]. This phenomenon is called interfacial activation. Various crystallographic structures of fungal [51, 52, 53, 54, 55] and bacterial [56, 57, 58, 59] lipases with and without ligands have shown that the interfacial activation can be at least partially attributed to a conformational change in these lipases. In fungal lipases, the interfacial activation is caused by the opening of an α-helical, amphipathic lid by a rotation around two hinge regions [39]. In the open, active form a significant hydrophobic surface is revealed. In bacterial lipases the conformational change involves rearrangements of multiple loops that surround the active site, and an increase in the solvent-exposed hydrophobic surface [58]. In the case of pancreatic lipases a more complex model including effects mediated by a colipase has been proposed [60, 61].

Recent studies on the activation of fungal and pancreatic lipases have indicated that the activation mechanism is controlled by a combination of hydrophobic and electrostatic interactions [62, 63] and is possibly influenced by the curvature of the lipid interface [60, 63]. The hydrophobic surface that is exposed in the open form of the lipase interacts with the lipid interface. Electrostatic interactions between the anionic groups of the lipid interface and the positively charged arginines in the hinge regions of the lid have been found to stabilize the open form [62, 63]. The importance of the electrostatic interactions in the activation of lipase is supported by experiments which show that lipase activation does not occur if the surface is zwitterionic [62, 63].

Although the importance of the hydrophobic and electrostatic interactions in the activation mechanism is now known, the exact mechanism of the interfacial activation remains unclear. In this thesis, computational methods were applied to address this problem.

1.3 A-kinase anchoring proteins

Many hormones and other biochemical signals function through receptors that trigger cell signaling pathways that lead to a phosphorylation or dephosphorylation of some key regulatory proteins inside the cell [64]. These signaling pathways are mediated by second messengers (Ca^{2+}, phospholipid, cAMP) that regulate the activity of the proteins responsible for the phosphorylation (kinases) and dephosphorylation (phosphoprotein phosphatases) events. cAMP activates cAMP-dependent protein kinase (PKA), which then phosphorylates its target proteins. Since PKA is able to phosphorylate a large number of different proteins, a mechanism is needed to ensure the selectivity of the phosphorylation events.

A-kinase anchoring proteins (AKAPs) are adapter proteins that are partially responsible for the specificity of the phosphorylation events catalyzed by PKA [65, 66, 67]. AKAPs do not have any catalytic activity themselves - their

role is to direct PKA and some other signaling proteins to certain intracellular structures and organelles. AKAPs are multi-domain proteins that typically consist of at least two domains: the PKA-binding domain and the targeting domain. The PKA-binding domain is an amphipathic α-helix that binds to the N-terminal part of the regulatory (R) subunit of PKA [68, 69]. AKAPs are specific towards the type of the regulatory subunit: RIα-specific, RIIα-specific and dual-specific AKAPs have been characterized [70, 71].

The targeting domain of the AKAPs anchors the PKA-AKAP complex to its correct subcellular location. AKAPs have been observed to bind to mitochondria, centrosomes, actin cytoskeleton, Golgi apparatus, microtubules, plasma membrane, vesicles, endoplasmic reticulum, dendrites, and nuclear membrane (reviewed in [65, 66, 72]). In addition to the PKA-binding domain and the targeting domain, some other types of domains have been observed in AKAPs. An example is AKAP79/150, which not only binds PKA, but also protein phosphatase PP2B [73] and protein kinase C [74]. Thus, this AKAP functions as a scaffolding protein that bridges together a multi-protein signaling complex [75].

Over 70 different AKAPs have been found to date [67]. Since the domain architecture of the AKAPs is not conserved, the domain structure and function of each AKAP has to be determined separately. For many of the currently known AKAPs, the domain architecture is not known or is poorly characterized, indicating that many more domains and functions remain to be discovered for the AKAPs. Sequence-based bioinformatics methods offer a powerful set of tools to characterize domain architectures, identify novel domains and predict novel functions for AKAPs. In this study, these methods were used to characterize the domain organization of an AKAP from *C. elegans*.

1.4 Aims of the present study

The specific aims of this thesis are:

- Characterize the local and global motions associated with the conformational change in *R. miehei* lipase and investigate how the dielectric of the medium affects these motions.
- Characterize the key electrostatic interactions in the *R. miehei* lipase family. Identify conserved sequence motifs, and investigate, whether the key residues identified by the electrostatic calculations are conserved.
- Investigate the domain architecture of A-kinase anchoring proteins (AKAPs). Identify novel domains, and make novel predictions for the AKAP function.

Chapter 2

Computational methods

2.1 Normal mode analysis

2.1.1 Introduction to normal mode analysis

Normal mode analysis (NMA) is a method used to calculate the low-frequency vibrational modes for a protein directly from the potential energy function by assuming that the potential energy surface is quadratic [29]. The advantage of NMA is that it yields a description of the global, elastic motions of proteins [29, 76, 77]. The disadvantage of NMA is that it samples only one energy minimum, assuming that the motion is harmonic when it in fact also contains anharmonic effects [78, 79, 80]. However, several studies have shown that despite its limitations, NMA is able to capture the biologically significant, collective motions in proteins [81, 82, 83]. These motions can be observed in several energy minima [79], indicating that certain motions are an inherent feature of a protein structure and not specific to a certain energy minimum. Further, in a study by Hayward *et al.* [84] on bovine pancreatic trypsin inhibitor a good qualitative correspondence was obtained between NMA and a principal component analysis of a molecular dynamics simulation, when the first principal component was excluded from the analysis. NMA has been successfully applied to characterize domain and other large-scale motions as well as conformational changes in several proteins [77, 81, 82, 83, 85, 86, 87, 88, 89, 90], investigate interactions between domains [91], explore energy landscapes [78], refine crystallographic *B*-factors [92, 93], and investigate the effects of point mutations on enzyme specificity [80]. In this thesis, an application of NMA to the characterization of the conformational change in *R. miehei* lipase is presented.

2.1.2 Theory of normal mode analysis

The vibrational motions of a molecule consisting of N atoms can be obtained by calculating the eigenvalues and eigenvectors of a $3N \times 3N$ mass-weighted Hessian matrix H [29, 76, 94]. The elements of the Hessian matrix are second

derivatives of the potential energy function, and can be expressed as

$$H_{ij} = \frac{\partial^2 V}{\partial x_i \partial x_j} \tag{2.1}$$

where V is the potential energy function and x_i are the Cartesian displacement coordinates. Assuming that the motions of the atoms in the molecule are harmonic, the equations of motion for the molecule reduce to the normal mode equation

$$Ha = \omega^2 Ma \tag{2.2}$$

where M is a $3N \times 3N$ diagonal matrix with masses of the atoms on the diagonal, ω is the angular frequency of the normal mode, and a is the vector of normal mode amplitudes.

Equation 2.2 can be rewritten as

$$Fq = \omega^2 q \tag{2.3}$$

where

$$F = M^{-1/2} H M^{-1/2} \tag{2.4}$$

is the force constant matrix in mass weighted coordinates and

$$q = M^{1/2} a \tag{2.5}$$

are the mass weighted normal mode amplitudes. The allowed values for ω^2 in Eq. 2.3, i.e. the eigenvalues of F, can be solved from the secular equation

$$|F - \omega^2 I| = 0 \tag{2.6}$$

where I is the identity matrix. This results in a set of $3N$ normal mode frequencies $\nu^{(k)}$ given by

$$\nu^{(k)} = \frac{\omega^{(k)}}{2\pi} \tag{2.7}$$

where $(\omega^{(k)})^2$ is the kth eigenvalue of matrix F. The eigenvectors $q^{(k)}$ can then be solved from

$$Fq^{(k)} = (\omega^{(k)})^2 q^{(k)} \tag{2.8}$$

2.1.3 Iterative procedures to obtain the normal modes

The frequencies $\nu^{(k)}$ are determined from the secular equation 2.6 by a diagonalization procedure. For large molecules like proteins that consist of thousands of atoms, direct diagonalization of the force constant matrix F is not computationally feasible due to the high computer memory demand. Since it is the collective, low-frequency vibrations that are usually of interest, it is appropriate to calculate only the lowest frequency modes rather than the whole spectrum. Several methods to compute only the low-frequency vibrations of proteins have been published, and have been reviewed by Perahia

and Mouawad [76]. One of the approaches that has been shown to yield the eigenvalues and eigenvectors without significant errors is the method of diagonalization in a mixed basis (DIMB) [29, 76]. DIMB uses an iterative procedure in which the approximate low-frequency modes are coupled with the Cartesian coordinates to obtain the exact low-frequency modes. In this study, DIMB was used to calculate the 47 lowest frequency vibrational modes of *R. miehei* lipase.

2.2 Electrostatic calculations

2.2.1 Introduction to electrostatic calculations

Electrostatic interactions have both local and global effects in proteins. Locally important electrostatic interactions include hydrogen bonds [95, 96], salt bridges [97, 98], and interactions that stabilize secondary structure elements [99, 100, 101]. The long-range electrostatic interactions play an important role e.g. in steering charged ligands towards enzyme active sites [102] and affecting formation of protein-protein complexes [103, 104].

Methods for investigating the role of electrostatic interactions in protein function and stability as well as in protein-ligand complexes have been developed [31, 35, 105, 106]. Brownian dynamics simulations have proven to be useful in understanding diffusion-controlled reactions and protein-protein association [107, 108, 109, 110]. Electrostatic potential surface calculations often reveal electrostatic patterns that are functionally important [30]. Calculations of electrostatic free energies, pK_a values of titrating groups, and titration curves have all been used extensively to investigate the function and pH-dependent stability of biomolecules and protein-ligand interactions [33, 36, 37, 38, 111, 112, 113, 114, 115, 116, 117]. In this thesis, pK_a calculations were applied to investigate the role of electrostatic interactions in the function and stability of the *R. miehei* lipase family.

The pK_a value is the pH at which an ionizable group is half-ionized. The groups that are usually treated as ionizable (titrating) in proteins include Asp, Glu, Lys, Arg and His. Computational methods to determine the pK_a values of the titrating groups have proven to be popular because the theoretically determined pK_a values can be compared to the pK_a values obtained by nuclear magnetic resonance spectroscopy [118, 119, 120, 121].

pK_a calculations are based on calculating the pK_a shifts relative to an isolated amino acid residue in solution. The pK_a values are affected mainly by two factors: exposure to solvent, and interactions with other charged and polar groups in the protein [37]. A charged residue that is buried in the protein typically has its ionization equilibrium shifted towards the neutral state. For acidic residues, the pK_a value is shifted upwards, and for basic residues, the pK_a value is shifted downwards. This effect is compensated to some extent by the attractive Coulombic interactions with other charged and polar residues in the protein.

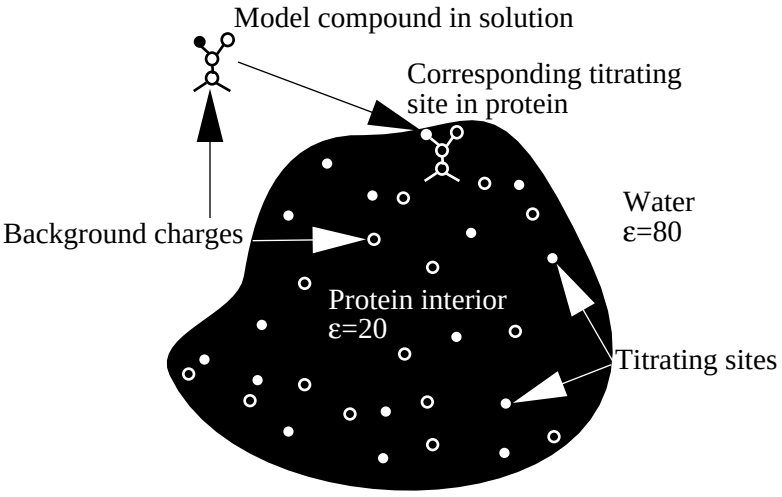

Figure 2.1: A two-dimensional representation of the protein model used in the pK_a calculations (adapted from [122]) . The protein is presented as a low dielectric continuum (ϵ=20) containing a number of background charges and titrating sites. The protein is embedded in a high dielectric medium (ϵ=80).

2.2.2 Poisson's equation

In electrostatic calculations, a protein is treated as a low-dielectric continuum containing a number of charges surrounded by a high-dielectric solvent continuum that contains mobile ions (Figure 2.1) [123]. Protein electrostatic calculations are based on solving the Poisson's equation

$$\nabla \cdot [\epsilon(r)\nabla\Phi(r)] = -4\pi\rho(r) \tag{2.9}$$

which relates the variation of the electrostatic potential $\Phi(r)$ in position r to the charge density distribution $\rho(r)$ in a medium that has a position-dependent permittivity (dielectric constant) of $\epsilon(r)$.

Since protein systems consist of two regions with different dielectric constants (protein and solvent), a modified version of the Poisson's equation must be used. Most methods available are based on numerical solutions of the linearized Poisson-Boltzmann (PB) equation

$$\nabla \cdot [\epsilon(r)\nabla\Phi(r)] - \bar{\kappa}^2(r)\Phi(r) = -4\pi\rho(r) \tag{2.10}$$

in which κ is the Debye-Hückel parameter and $\kappa^2 = 8\pi Ne^2 I/1000DkT$, where N is Avogadro's number, e is the electronic charge, I is the ionic strength, k is Boltzmann's constant and T is the absolute temperature. In equation 2.10 a modified Debye-Hückel parameter $\bar{\kappa} = D^{1/2}\kappa$ is used.

In this thesis, a numerical technique based on a finite-difference (FD) method was used to solve the linearized PB equation within the program UHBD (University of Houston Brownian dynamics) [31, 35, 124]. In the FD method the protein is placed at the center of a three-dimensional grid, and a system of linear equations is used to represent the potential at each grid

point as a function of the potential at surrounding grid points. The system of the linear equations is then solved iteratively using a preconditioned conjugate gradient algorithm.

The accuracy of the FDPB-calculations is dependent on the spacing of the grid where the protein is placed. The focussing [106] method aims at improving the accuracy of the FDPB-calculations by performing a series of FD calculations, starting by calculating the potential for the whole molecule using a large grid spacing, and then narrowing down by using the boundary conditions from the previous run and a smaller grid spacing.

2.2.3 Calculation of the pK_a values

The pK_a values of the titrating groups in a protein are calculated by starting with the model compound solution pK_a values (p$K_{a,model}$) [34, 123]. The calculation is based on the idea that the group is moved from solution to its specific location in the protein, and the electrostatic effects caused by the move are calculated. The energetics of an ionizable group is affected by three electrostatic terms [34, 122, 125, 126]. The first term is the Born or desolvation energy, which is the energy penalty caused by removing the group from the solvent. The second term, the "background" term, is the interaction energy of the group with the background partial charges in a neutral protein. The third term accounts for the interaction of the group with all other ionized groups in the protein. By taking the first two energy terms into account, the intrinsic pK_a value (p$K_{a,int}$) of an ionizing group in a hypothetically neutral protein can be calculated by

$$\mathrm{p}K_{\mathrm{a,int}}(i) = \mathrm{p}K_{\mathrm{a,model}}(i) - z_i \frac{\Delta\Delta G_i}{2.303RT} \tag{2.11}$$

where z_i is the charge at site i in units of the electronic charge (for protein side chains the allowed values are -1, 0 and +1), $\Delta\Delta G_i$ is the difference in the ionization free energies of the residue in the folded but uncharged protein and isolated in solution, and R is gas constant. For each titrating site, two FDPB calculations are required to obtain $\Delta\Delta G_i$: one for calculating the ionization free energy of the residue isolated in solution, and the other one in the folded but uncharged protein. The ionization is modeled by adding a unit charge to the charge of the atom which becomes protonated or deprotonated.

The actual pK_a value for a site i can be calculated from p$K_{a,int}(i)$ using the Tanford-Roxby method [127]. According to this method, the actual pK_a value is obtained from p$K_{a,int}$ by subtracting the term that represents the pairwise electrostatic interaction between the site and every other site in the protein.

$$\mathrm{p}K_{\mathrm{a}}(i) = \mathrm{p}K_{\mathrm{a,int}}(i) - \sum_{j \neq 1} \frac{W_{ij}}{2.303 z_i kT} \tag{2.12}$$

where W_{ij} is the pairwise interaction free energy between i and j. For n titratable sites, 2^n different states must be enumerated in the Tanford and

Table 2.1: References and WWW-addresses of the databases used in this thesis.

Database	Ref.	WWW
GenBank	[129]	http://www.ncbi.nlm.nih.gov/
SWISS-PROT	[130]	http://www.expasy.ch/sprot/sprot-top.html
Protein Data Bank	[131]	http://www.rcsb.org/pdb/
FSSP	[132]	http://www2.ebi.ac.uk/dali/fssp/
SCOP	[133]	http://scop.berkeley.edu/
InterPro	[134]	http://www.ebi.ac.uk/interpro/index.html
SMART	[135]	http://smart.embl-heidelberg.de/

Roxby method. UHBD uses the cluster method developed by Gilson [128] which reduces the computational demands by clustering the titrating sites, solving the ionization polynomial exactly within the cluster and using a mean field approximation to treat intercluster interactions.

2.3 Bioinformatics methods

The bioinformatics methods used in this thesis include gene structure predictions, searching DNA and protein sequence and structure databases as well as motif and domain databases, generating pairwise and multiple sequence alignments, determining conserved sequence motifs in a set of protein sequences, generating motif and profile hidden Markov models and secondary structure predictions. The WWW- addresses of the databases used in this thesis are listed in Table 2.1, and the bioinformatics methods are listed in Table 2.2.

Several databases were searched to find protein and DNA sequences and structures, and to identify functional sites and domains. Sequence and structure databases used in this thesis include GenBank [129], SWISS-PROT [130] and Protein Data Bank (PDB) [131]. FSSP [132] was used to identify structural neighbors and to obtain structural alignments, and the SCOP (Structural Classification of Proteins) database [133] was used to define structural protein families and subfamilies. InterPro [134] is an integrated database that contains the following databases: (1) PROSITE [136] (biologically significant motifs and profiles in protein sequences), (2) Pfam [137] (multiple sequence alignments and hidden Markov models of protein domain families), (3) PRINTS [138] (protein family fingerprints, groups of aligned protein sequence motifs) and (4) Blocks [139] (ungapped multiple alignments of conserved regions in proteins). SMART is a database that contains well-curated, gapped sequence alignments of domains that occur in signaling proteins [135].

PSI-BLAST [140] was used to identify sequence homologues by searching the NCBI (National Center for Biotechnology Information) non-redundant protein sequence database, which contains a non-redundant set of protein sequences retrieved from the GenBank CDS (coding sequence) translations,

Table 2.2: References and WWW-addresses of the bioinformatics programs used in this thesis.

Program	Ref.	WWW
PSI-BLAST	[140]	http://www.ncbi.nlm.nih.gov/BLAST/
Pairwise BLAST	[140, 142]	http://www.ncbi.nlm.nih.gov/BLAST/
ClustalW	[143, 144]	http://www.ebi.ac.uk/clustalw/
HMMer	[145]	http://hmmer.wustl.edu/
TEXshade	[146]	http://homepages.uni-tuebingen.de/beitz/txe.html
BOXSHADE	[147]	http://www.ch.embnet.org/software/BOX_form.html
MEME	[148]	http://meme.sdsc.edu/meme/website/
GENSCAN	[3]	http://CCR-081.mit.edu/GENSCAN.html
JPred^2	[15]	http://jura.ebi.ac.uk:8888//
PHD	[16]	http://maple.bioc.columbia.edu/predict-protein/

SWISS-PROT, PDB, PIR [141] (Protein Information Resource) and PRF (Protein Research Foundation). Several iterative searches were performed using PSI-BLAST, and the significant hits found in each run were used to construct a new position-specific score matrix for the next run. Pairwise BLAST [140, 142] was used to identify local similarities in protein sequences.

Sequence alignments were generated using four different programs: ClustalW, pairwise BLAST, FSSP and HMMer. ClustalW [143, 144] generates multiple sequence alignments progressively, starting with the most similar sequences and adding the more distant ones gradually. Pairwise BLAST was used to generate local sequence alignments. Structural alignments were obtained from FSSP. HMMer [145] was used to build a profile hidden Markov model of a set of structurally aligned sequences and to align other sequences to the model. TEXshade [146] and BOXSHADE [147] were used to shade the alignments.

Hidden Markov models (HMMs) were constructed for certain domains and sequence motifs, and these models were used to search for additional occurrences of these motifs. HMMer was used to build profile HMMs, and MetaMEME [149] was used to construct motif HMMs from sequence motifs generated by MEME (Multiple Expectation-maximization for Motif Elicitation) [148]. The difference between a profile HMM and a motif HMM is that the motif model is based on modeling ungapped blocks of sequence motifs whereas the profile model allows insertions and deletions in the target sequence [150].

Other methods used in the thesis include gene and secondary structure predictions. Gene structure predictions were carried out using GENSCAN [3].

Secondary structure predictions were done using JPred^2 [15] and PHD [16]. JPred^2 combines predictions from six different secondary structure prediction algorithms that all use evolutionary information obtained from homologous sequences [15].

Chapter 3

Results and discussion

3.1 Computational studies of lipase structure and function

3.1.1 Structural features of the *R. miehei* lipase family

Full coordinates for three-dimensional structures of three *R. miehei* lipase (RmL) family members can be found in the PDB: closed and open conformations of RmL (PDB ID's 3tgl and 4tgl), a closed conformation of *Thermomyces* (*Humicola*) *lanuginosus* lipase (1tib) and an intermediate conformation of the *Rhizopus niveus* lipase (RnL, 1lgy) (Figure 3.1). The coordinate file of the *Penicillium camembertii* lipase structure (1tia) contains only the Cα-coordinates, and was not used in this study. *Rhizopus oryzae* lipase has an identical sequence and structure to RnL, and was also left out of the study. The fairly small RMS-deviations of the Cα-atoms between the different lipases (1.1-1.9 Å) indicate that lipases belonging to the RmL family have only minor differences in their structures.

Figure 3.1 shows that proteins belonging to the RmL family have an α/β secondary structure in which the five central, parallel β-strands are surrounded by one antiparallel strand on the C-terminal side and by two antiparallel strands on the N-terminal side. The strands form an eight-strand singly wound β-sheet. The β-sheet is surrounded by three α-helices on the proximal side, one long N-terminal α-helix on the distal side, and one α-helix covering the active site. In the RmL family, the active site is located in a shallow groove covered by the α-helical lid in the inactive, closed conformation. At the interface between lipid and water, a conformational change is observed, in which the lid opens up by rotating around two hinge regions [51]. Figure 3.1 shows that the catalytic serine is located in a tight β-ϵSer-α turn that is formed by a Gly-X-Ser-X-Gly-motif characteristic of lipases [151]. In the active site, the catalytic triad residues Ser, Asp and His are bonded to each other by a network of hydrogen bonds.

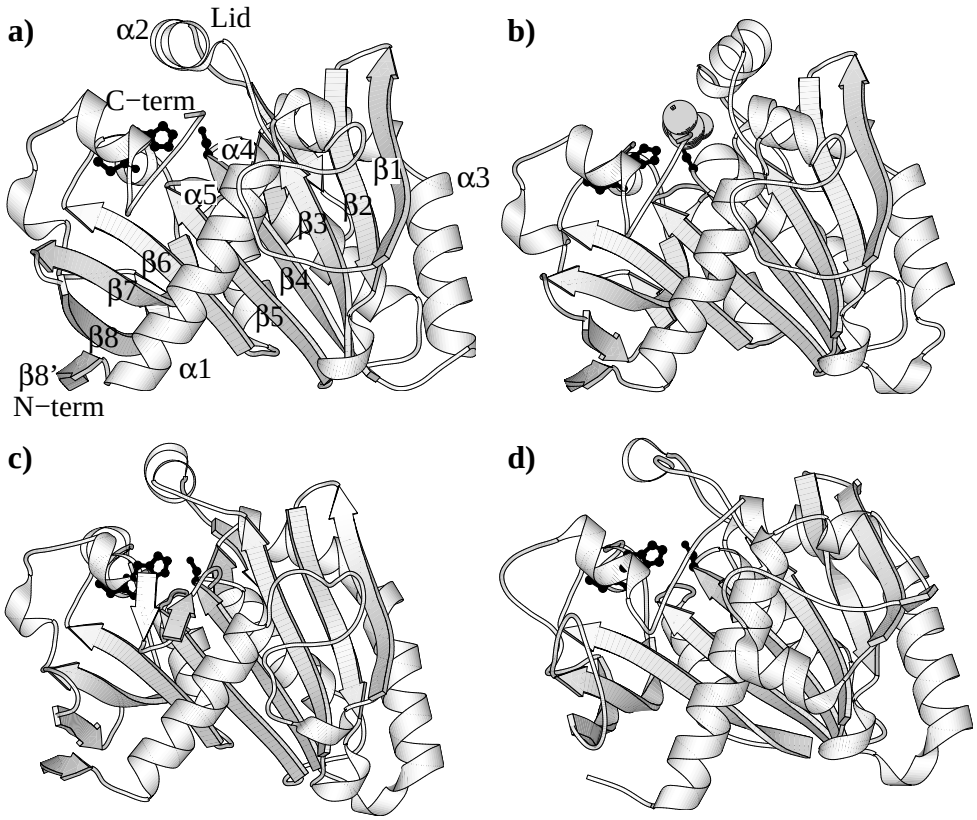

Figure 3.1: Schematic representations of the a) closed and b) open conformations of *R. miehei* lipase, c) an intermediate conformation of *R. niveus* lipase and d) a closed conformation of *T. lanuginosus* lipase. Side chains of the catalytic residues are shown as ball-and-stick. In b), an inhibitor bound to the active site is shown.

3.1.2 Normal mode analysis (Paper I)

Normal mode analysis (NMA) was performed for the closed and open conformations of RmL to characterize the local and global motions associated with the conformational change. Visual inspection of the normal mode trajectories and the scalar products between the reaction coordinate and the Cα eigenvectors suggest that the conformational change is well described by the low-frequency normal modes.

Normal mode calculations were performed in both the low and high dielectric media to simulate the difference between water and the lipid interface. These calculations demonstrate that fluctuations of the first hinge of the mobile lid region increase when the lipase is moved from a high to a low dielectric medium. This result suggests that the first hinge may be involved in the initiation phase of the conformational change.

The role of the positively charged Arg86 in the lid of RmL was investigated. When Arg86 was neutralized, normal modes of the closed RmL in a high dielectric medium showed a deformation of the lid α-helix through increased

fluctuation of the second hinge. This indicates that Arg86 may have a role in the intrinsic stabilization of the lid α-helix.

The low-frequency normal modes reveal a global breathing motion as well as several local loop motions that are coupled with the lid displacement. Breathing motions in proteins usually involve flexing of secondary structure elements or concerted motions of domains or subdomains, and are often associated with ligand binding. A classic example of a breathing motion is myoglobin, in which a breathing motion facilitates oxygen binding and release [152, 153]. Recent studies suggest that breathing motions could also be involved e.g. in ion channel gating [154] and in the function of human glutathione transferase P1-1 [155]. The magnitude of the breathing motion observed in RmL in this study is very small; the radius of gyration changes only about 0.1% during the breathing motion. However, it is known that NMA gives more a qualitative rather than a quantitative description of the motions, and it often underpredicts their magnitude [84].

The conformational change in lipases has been studied extensively using theoretical simulation methods, but experimental data is scarce. Several experimental methods to investigate conformational changes in proteins do exist, and include e.g. nuclear magnetic resonance (NMR) spectroscopy [156], time-resolved X-ray crystallography [157], site-direct spin labeling [158], circular dichroism and synchrotron radiation circular dichroism spectroscopy [159], single pair fluorescence resonance energy transfer (spFRET)[160], atomic force microscopy [161, 162], and cryo-electron microscopy [163]. Of these methods, NMR has been used to identify structural changes associated with the interfacial activation in phospholipase A_2 [164, 165]. Homonuclear 2D NMR has been applied to investigate the interaction of mammalian colipases with bile salt micelles [166].

We observe that one of the potential methods to investigate further the conformational change in lipases is spFRET. FRET is based on an idea that the donor fluorophore's fluorescence intensity and excited state lifetime and the acceptor fluorophore's emission intensity are dependent on the distance between the donor and the acceptor [167]. FRET is commonly used to measure distances between 20 and 60 Å, and it has been successfully used to characterize the conformational change for instance in myosin upon ATP binding and hydrolysis [168]. For myosin, distance differences of only a few Ångströms could be measured, indicating that this method could be used to characterize the lid opening as well as the breathing motion observed in RmL by normal mode analysis.

3.1.3 Electrostatic calculations (Papers II and III)

In this thesis, continuum electrostatic calculations were used to identify the key electrostatic interactions that are important for the stability, enzymatic function, and for the activation mechanism of the RmL family of proteins.

In Paper II, the residues, whose electrostatic interaction energy with the lid

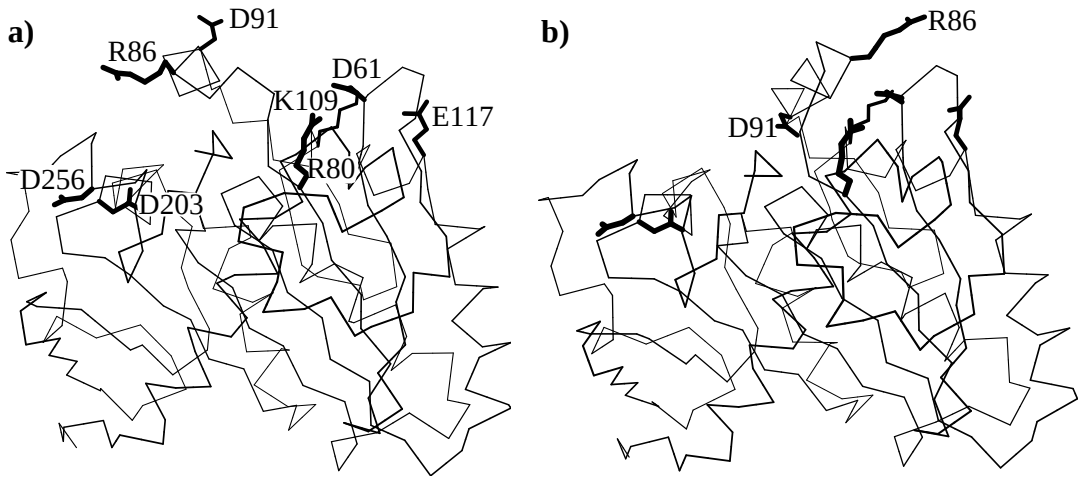

Figure 3.2: Cα-traces of the a) closed and b) open conformations of *R. miehei* lipase. The two charged lid residues (Arg86 and Asp91) as well as residues, whose electrostatic interaction energy with the lid is most perturbed by the change of the solvent dielectric, are labeled.

region is most perturbed by the change of the solvent dielectric, were identified in RmL (Figure 3.2). By changing the solvent dielectric, an attempt was made to model the change of the lipase environment from water to lipid and vice versa, and to determine which interactions may affect the conformational change.

In Paper III, a network of electrostatic interactions was characterized that connects the lipase active site residues to the mobile lid region (Figure 3.3). This network of interactions was shown to be well conserved in 12 RmL homologues, suggesting that it has either structural or functional importance. Three of the network members, Arg80, Asp91 and Asp203, were also identified as "key electrostatic residues" in Paper II.

Most of the electrostatic studies carried out to date have focused on investigating the pairwise interactions in proteins. The concept of electrostatic networking, in which more than two residues form a functionally important network of electrostatic interactions, is fairly new. A few interesting studies on the role of electrostatic networking in proteins have been published. In a paper by Sheinerman *et al.* [104] the authors note that networks formed by ionic and hydrogen bonding groups may play an important role in stabilizing protein-protein complexes. They use a bifurcated ion pair in the barstar-barnase complex as an example. The bifurcated ion pair consist of Asp39 in barstar and Arg83 and Arg87 in barnase. The authors point out that although either of these two salt bridges would be unstable in isolation, the bifurcated ion pair is energetically favorable. The same authors have investigated interfacial salt bridges in three protein-protein complexes and conclude that interfacial salt bridges are stabilized by other ionic or hydrogen bonding interactions that form networks of electrostatic interactions. Further, several

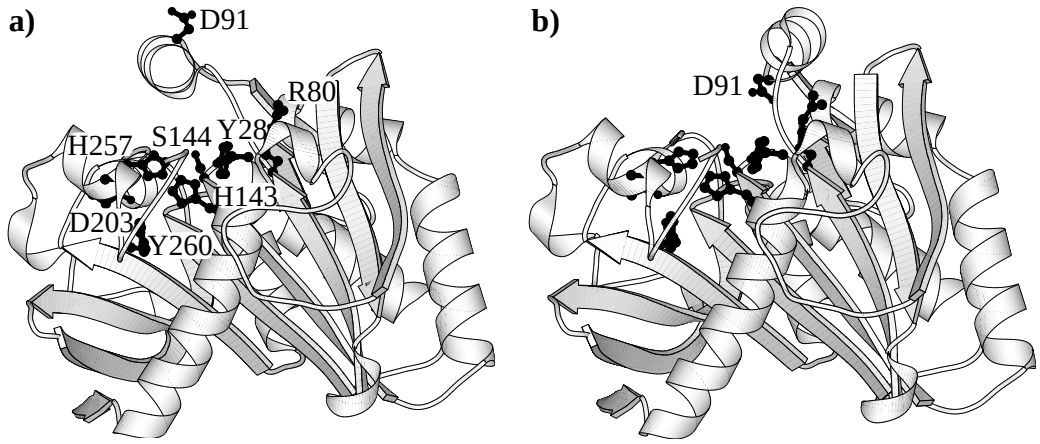

Figure 3.3: Schematic representations of the a) closed and b) open conformations of *R. miehei* lipase. Residues forming the electrostatic network are labeled.

theoretical and experimental studies have suggested that an increased number of ion pairs and ion pair networks may enhance the thermal stability of hyperthermophilic proteins [169, 170, 171, 172, 173, 174].

An example of the role of hydrogen bonds in protein structure and function is phospholipase A_2 (PLA2), in which a hydrogen bonding network connects the active site to the interfacial recognition site [165, 175, 176]. This network of hydrogen bonds is interesting, especially when taken into account that the fungal lipase family and the PLA2 family do belong to different fold classes. According to the SCOP database, the fungal lipase family is part of the α/β class, whereas the PLA2 family belongs to the all-α class. We identify two fundamental differences between the hydrogen bonding network in PLA2 and the electrostatic network in the RmL family. First, the electrostatic network in the RmL family contains two residues (Arg80 and Asp91) that are not directly hydrogen bonded to any other network residues. Second, on the basis of mutation studies performed for PLA2 and the RmL family, it appears that these two networks may serve different purposes. Yuan *et al.* [165] conclude that the function of the hydrogen bonding network in PLA2 is structural rather than functional, since mutations in this network do not necessarily lead to a significant suppression of the enzymatic activity, whereas mutations of Tyr28, His143 and Asp91 in *Rhizopus oryzae* lipase have been observed to lead to a significant loss of activity [177]. Thus, mutations in residues belonging to the electrostatic network do affect the activity of the RmL family members.

The idea of electrostatic networks that connect functionally important regions in proteins would be worth a further study. A large-scale study of several protein families using continuum electrostatics methods combined with sequence motif methods would reveal whether electrostatic networking is a common feature of protein structure. The study could be combined with a mutation study in which residues identified as being part of an electrostatic

network are mutated, and the effects of the mutations on the stability, activity and kinetics of the protein are investigated.

3.2 Identification of the domain architecture in the A-kinase anchoring protein from *C. elegans* (Paper IV)

The domain architecture of the *C. elegans* AKAP ($AKAP_{CE}$) was investigated using sequence-based bioinformatics methods. $AKAP_{CE}$ was shown to share two domains with SARA (Smad anchor for receptor activation). SARA functions as an adapter protein in the TGFβ (transforming growth factor β) signaling pathway by anchoring receptor-regulated Smads (Smad2 and Smad3) to the TGFβ receptor [178](Figure 3.4). The TGFβ signaling pathway has an important function in cell growth and differentiation, and mutations disabling components in this pathway have been shown to be associated with certain types of cancers [179].

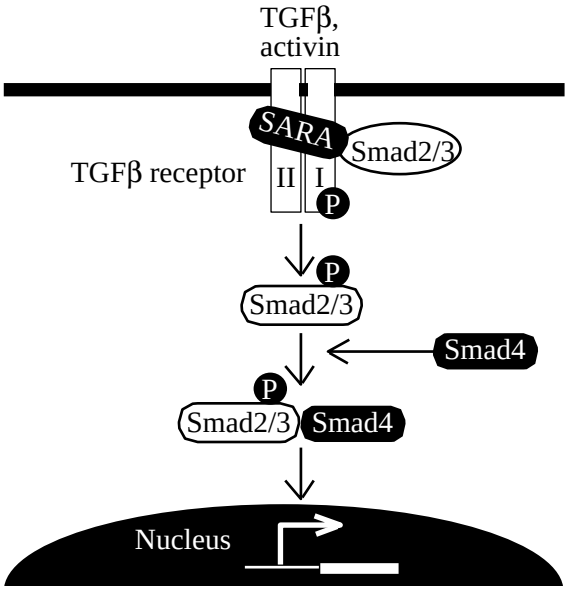

Figure 3.4: A model for the TGFβ signaling pathway (adapted from [180]). Ligand binding to the TGFβ receptor leads to a formation of a heteromeric receptor complex that contains both receptor II and receptor I [181]. Receptor II activates receptor I by phosphorylating it. Receptor I then phosphorylates a receptor-regulated Smad (Smad2/3), which leads to a conformational change and dissociation of the Smad from the receptor. The dissociated Smad forms a complex with a common mediator Smad (Smad4), and the complex is translocated to the nucleus, where it affects transcription of certain genes. SARA functions in the pathway as an adapter protein by anchoring Smad2/3 to the receptor.

The two domains that $AKAP_{CE}$ was shown to share with SARA are the FYVE-domain and the TGFβ receptor binding domain (Figure 3.5). Of these two domains, the FYVE-domain has been characterized earlier, and it is known that FYVE-domain containing proteins localize to membranes that contain phosphatidylinositol-3-phosphate [182, 183, 184]. The TGFβ receptor binding domain has not been characterized earlier. All that is known about this domain is that it interacts with the TGFβ receptor [178] - hence we call this domain the TGFβ receptor binding domain. In this study, two conserved regions as well as several conserved residues and sequence motifs were identified in the TGFβ receptor binding domain.

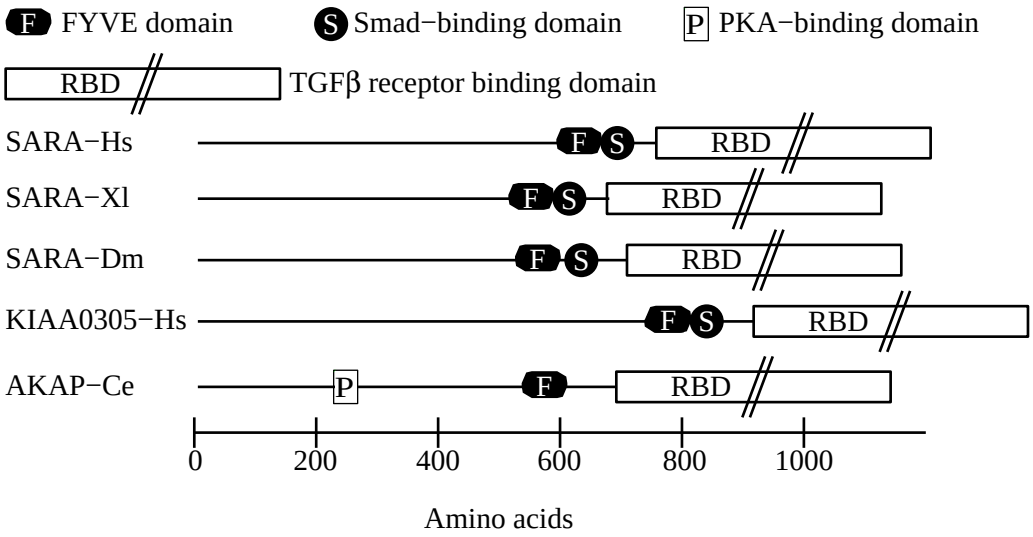

Figure 3.5: Domain architecture of the $AKAP_{CE}$ homologues. Sequence accession numbers (GenPept or Swiss-Prot/TrEMBL): SARA-Hs (NP_004790), SARA-Xl (AAC99463), SARA-Dm (AAF64468), KIAA0305-Hs (BAA20764), AKAP-Ce (Q18930). Abbreviations of the species used: Hs, *Homo sapiens*; Xl, *Xenopus laevis*; Dm, *Drosophila melanogaster*; Ce, *Caenorhabditis elegans*. The presence of a Smad-binding domain in KIAA0305 is uncertain. In this figure, the TGFβ receptor binding domain has been truncated.

Because $AKAP_{CE}$ was shown to share not only the FYVE-domain, but also the TGFβ receptor binding domain with SARA, a novel AKAP function was proposed for $AKAP_{CE}$ as a TGFβ receptor binding protein. This is supported by recent experiments, which show that three TGFβ family members stimulate PKA in the cell [185, 186, 187]. Although evidence exists for the interaction of PKA with the TGFβ signaling pathway, the mechanism of this interaction has not been understood earlier. Our hypothesis is that $AKAP_{CE}$ anchors PKA to the receptor, thus facilitating the interaction between PKA and the pathway.

Signaling pathways in *C. elegans* and in human have been observed to be fairly similar, and many components of the human TGFβ signaling pathway have been identified also in *C. elegans* [188]. Thus, one could expect that a SARA-orthologue could possibly exist in *C. elegans*. We find that $AKAP_{CE}$ is

the only SARA-homologue in the *C. elegans* genome. This raises the question, whether SARA is actually an AKAP and binds PKA. Our attempt to identify a putative PKA-binding site in SARA using a hidden Markov model did not yield any definitive answers. We conclude that the fact that no putative PKA-binding site was identified in SARA does not mean that SARA would not bind PKA. The reason for this is that the PKA-binding domain is fairly short and has only a few conserved residues, as can be seen in Figure 3.6.

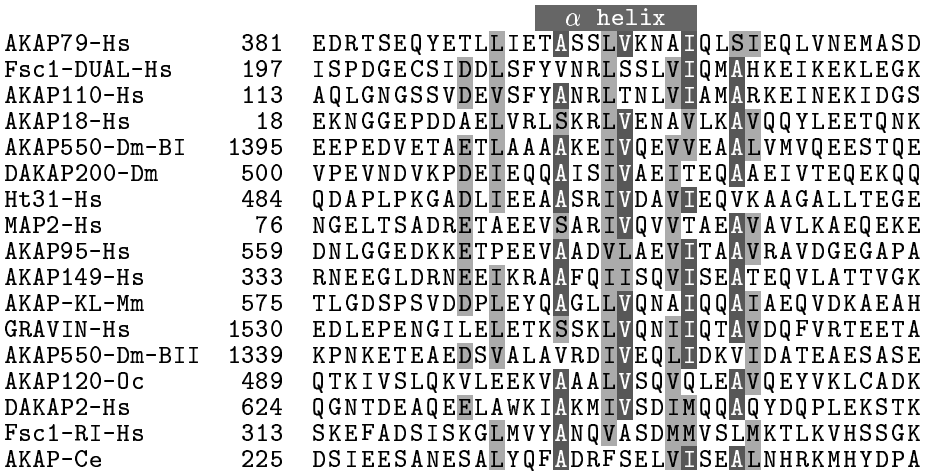

Figure 3.6: A sequence alignment of the PKA-binding region that was used to construct a hidden Markov model in order to search for a putative PKA-binding domain in SARA. The approximate location of the PKA-binding α-helix is labeled above the alignment. Ten residues on each side of the PKA-binding domain were included in the alignment. The following AKAPs were used in the model: AKAP79-Hs (Swiss-Prot/TrEMBL accession number P24588), Fsc1-Hs (GenPept accession number AAC79433), AKAP110-Hs (O75969), AKAP18-Hs (O43687), AKAP550-Dm (O16024), DAKAP200-Dm (AAD47200), Ht31-Hs (Q14572), MAP2-Hs (P11137), AKAP95-Hs (O43823), AKAP149-Hs (Q92667), AKAP-KL-Mm (O54931), GRAVIN-Hs (AAC51366), AKAP120-Oc (Q28628), DAKAP2-Hs (AAB92260) and AKAP-Ce (Q18930). Abbreviations of the species used: Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; Mm, *Mus musculus*; Oc, *Oryctolagus cuniculus*; Ce, *Caenorhabditis elegans*. Conserved identical residues are shaded in dark gray and similar residues in light gray.

AKAPs are adapter proteins, and some of them are known to function as scaffolding proteins that bridge together multi-protein signaling complexes [75]. It is known that adapter and scaffolding proteins do have a role in controlling specificity and selectivity within signaling pathways by assembling correct pools of cell signaling components together [189]. In addition, certain evidence exists that some adapter and scaffolding proteins may also function as 'hubs' that connect two signaling pathways and enable their interaction [190]. An interesting example of an AKAP that functions as a 'hub' is AKAP79/150 [75].

This AKAP recruits PKA, protein kinase C and protein phosphatase PP2B to the β_2-adrenergic receptor, which enhances phosphorylation of the receptor and facilitates activation of the MAP kinase pathway. The novel AKAP function presented in this thesis indicates that $AKAP_{CE}$ could also function as a 'hub' that facilitates interaction between PKA and the TGFβ signaling pathway by recruiting PKA to the TGFβ receptor.

Chapter 4

Conclusions and future perspectives

In this thesis, computational methods were applied to investigate protein sequence and structure and their relationship to protein function. Efforts were made in three areas. First, the local and global motions associated with the conformational change in *R. miehei* lipase (RmL) were characterized using normal mode analysis. Second, the key electrostatic interactions in the RmL family were identified by solving numerically the linearized Poisson-Boltzmann equation. Third, the domain architecture of an AKAP from *C. elegans* (AKAP_{CE}) was investigated using sequence-based bioinformatics methods.

The lipase structure-function relationship was studied using two complementary methods: normal mode analysis and electrostatic calculations. In normal mode analysis, the collective (and often biologically significant) motions of a protein are obtained directly from the potential energy function, whereas the electrostatic calculations give a detailed description of the role of electrostatic interactions in the structure, function and stability of proteins. The normal modes presented in this thesis reveal, how local loop motions as well as a global breathing motion are coupled with the conformational change in RmL. The normal modes were observed not only to describe the structural plasticity of the lipase molecule but also to indicate how the motions are associated with the function (in this case, conformational change).

The electrostatic calculations performed for the RmL family yield two significant results. First, the residues, whose electrostatic interaction energy with the lid region is most perturbed by the change of the solvent dielectric, were identified in RmL. Second, a network of electrostatic interactions was discovered in the RmL family. This network is an important feature of the lipase structure, since it connects the catalytic triad to the mobile lid region. The electrostatic calculations were combined with sequence motif methods to show that the network residues are well conserved in the RmL family and their homologues, and that most of these residues are located in sequence motifs. Since the network residues were observed to be conserved in evolution, they can be

assumed to have either a functionally or structurally important role.

The sequence-function relationship of $AKAP_{CE}$ was investigated using bioinformatics methods. In this study, only sequence-based methods were used, because very little structural information about AKAPs is available. Two significant findings were made. First, a novel domain (the TGFβ receptor binding domain) was characterized, and two highly conserved regions as well as several conserved sequence motifs were identified in the five homologues that were found to contain this domain. Second, the domain architecture of the $AKAP_{CE}$ led us to propose a novel AKAP function as a TGFβ receptor binding protein. From the point of view of the protein sequence-structure-function paradigm, the AKAP study shows how protein sequence analysis can be applied to identify motifs and domains and how it can be used in the functional annotation of proteins.

The studies presented in this thesis demonstrate that, by combining sequence-level bioinformatics methods with structural information, an understanding of the sequence-structure-function relationship of whole protein families can be obtained. In the future, the emphasis will shift from single-molecule studies to studies of complete protein families and large protein assemblies as well as metabolic and signaling pathways. This shift will be enabled by the current genome sequencing and structural genomics projects, as well as other high-throughput analysis methods such as functional genomics and proteomics methods.

Bibliography

[1] R.F. Doolittle. On the trail of protein sequences. *Bioinformatics*, 16:24–33, 2000.

[2] E.W. Myers, G.G. Sutton, A.L. Delcher, I.M. Dew, D.P. Fasulo, M.J. Flanigan, S.A. Kravitz, C.M. Mobarry, K.H.J. Reinert, and K.A. Remington *et al.* A whole-genome assembly of *Drosophila*. *Science*, 287:2196–2204, 2000.

[3] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[4] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, and R.F. Galle *et al.* The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195, 2000.

[5] A.M. Lesk and C. Chothia. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, 136:225–270, 1980.

[6] C.P. Ponting, J. Schultz, R.R. Copley, M.A. Andrade, and P. Bork. Evolution of domain families. *Adv. Protein Chem.*, 54:185–244, 2000.

[7] A. Bateman and E. Birney. Searching databases to find protein domain organization. *Adv. Protein Chem.*, 54:137–157, 2000.

[8] R.F. Doolittle. Similar amino acid sequences: chance or common ancestry? *Science*, 214:149–159, 1981.

[9] D.G. Higgins, J.D. Thompson, and T.J. Gibson. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, 266:383–402, 1996.

[10] L. Aravind, K.S. Makarova, and E.V. Koonin. Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, 28:3417–3432, 2000.

[11] D.G. Higgins. Amino acid-based phylogeny and alignment. *Adv. Protein Chem.*, 54:99–135, 2000.

[12] C.M. Dobson and M. Karplus. The fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Struct. Biol.*, 9:92–101, 1999.

[13] B. Honig. Protein folding: from the Levinthal paradox to structure prediction. *J. Mol. Biol.*, 293:283–293, 1999.

[14] A.R. Dinner, A. Šali, L.J. Smith, C.M. Dobson, and M. Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, 25:331–339, 2000.

[15] J.A. Cuff, M.E. Clamp, A.S. Siddiqui, M. Finlay, and G.J. Barton. JPred: a consensus secondary structure prediction server. *Bioinformatics*, 14:892–893, 1998.

[16] B. Rost. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.*, 266:525–539, 1996.

[17] T.L. Blundell, B.L. Sibanda, M.J.E. Sternberg, and J.M. Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347–352, 1987.

[18] M.S. Johnson, N. Srinivasan, R. Sowdhamini, and T.L. Blundell. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.*, 29:1–68, 1994.

[19] D.T. Jones, W.R. Taylor, and J.M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.

[20] D. Jones and J. Thornton. Protein fold recognition. *J. Comput. Aided. Mol. Des.*, 7:439–456, 1993.

[21] C.A. Orengo, J.E. Bray, T. Hubbard, L. LoConte, and I. Sillitoe. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*, Suppl 3:149–170, 1999.

[22] J.M. Thornton, C.A. Orengo, A.E. Todd, and F.M.G. Pearl. Protein folds, functions and evolution. *J. Mol. Biol.*, 293:333–342, 1999.

[23] L. Lo Conte, B. Ailey, T.J.P. Hubbard, S.E. Brenner, A.G. Murzin, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, 28:257–259, 2000.

[24] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[25] B. Lesyng and J.A. McCammon. Molecular modeling methods. Basic techniques and challenging problems. *Pharmacol. Ther.*, 60:149–167, 1993.

[26] J.M. Haile. *Molecular dynamics simulation: elementary methods.* John Wiley & Sons, New York, 1992.

[27] R.C. Wade, M.E. Davis, B.A. Luty, J.D. Madura, and J.A. McCammon. Gating of the active site of triose phosphate isomerase: Brownian dynamics simulations of flexible peptide loops in the enzyme. *Biophys. J.*, 64:9–15, 1993.

[28] R.C. Wade, B.A. Luty, E. Demchuk, J.D. Madura, M.E. Davis, J.M. Briggs, and J.A. McCammon. Simulation of enzyme-substrate encounter with gated active sites. *Nat. Struct. Biol.*, 1:65–69, 1994.

[29] L. Mouawad and D. Perahia. Diagonalization in a mixed basis: a method to compute low-frequency normal modes for large macromolecules. *Biopolymers*, 33:599–611, 1993.

[30] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268:1144–1149, 1995.

[31] J.D. Madura, J.M. Briggs, R.C. Wade, M.E. Davis, B.A. Luty, A. Ilin, J. Antosiewicz, M.K. Gilson, B. Bagheri, and L. R. Scott *et al.* Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.*, 91:57–95, 1995.

[32] M.E. Davis and J.A. McCammon. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.*, 90:509–521, 1990.

[33] D. Bashford and K. Gerwert. Electrostatic calculations of the p$K_{\rm a}$ values of ionizable groups in bacteriorhodopsin. *J. Mol. Biol.*, 224:473–486, 1992.

[34] J. Antosiewicz, J.A. McCammon, and M.K. Gilson. Prediction of pH-dependent properties of proteins. *J. Mol. Biol.*, 238:415–436, 1994.

[35] J.D. Madura, M.E. Davis, M.K. Gilson, R.C. Wade, B.A. Luty, and J.A. McCammon. Biological applications of electrostatic calculations and Brownian dynamics simulations. In K.B. Lipkowitz and D.B. Boyd, editors, *Reviews in Computational Chemistry*, volume V, pages 229–267. VCH Publishers, New York, 1994.

[36] V. Dillet, H.J. Dyson, and D. Bashford. Calculations of electrostatic interactions and p$K_{\rm a}$s in the active site of *Escherichia coli* thioredoxin. *Biochemistry*, 37:10298–10306, 1998.

[37] A.D. Elcock and J.A. McCammon. Electrostatic contributions to the stability of halophilic proteins. *J. Mol. Biol.*, 280:731–748, 1998.

[38] H.W.T. van Vlijmen, S. Curry, M. Schaefer, and M. Karplus. Titration calculations of foot-and-mouth disease virus capsids and their stabilities as a function of pH. *J. Mol. Biol.*, 275:295–308, 1998.

[39] Z.S. Derewenda. Structure and function of lipases. *Adv. Protein Chem.*, 45:1–52, 1994.

[40] P. Desnuelle. Pancreatic lipase and phospholipase. In P. Desnuelle, H. Sjöström, and O. Norén, editors, *Molecular and Cellular Basis of Digestion*, pages 275–296. Elsevier, Amsterdam, 1986.

[41] F.R. Hassanien and K.D. Mukherjee. Isolation of lipase from germinating oilseeds for biotechnological processes. *J. Am. Oil Chem. Soc.*, 63:893–897, 1986.

[42] E. Lie, A. Persson, and G. Molin. Screening for lipase-producing microorganisms with a continuous cultivation system. *Appl. Microbiol. Biotechnol.*, 35:19–20, 1991.

[43] E.N. Vulfson. Industrial applications of lipases. In P. Woolley and S.B. Petersen, editors, *Lipases - their structure, biochemistry and application*, pages 271–288. Cambridge University Press, Cambridge, 1994.

[44] K.-E. Jaeger and M.T. Reetz. Microbial lipases form versatile tools for biotechnology. *Trends Biotechnol.*, 16:396–403, 1998.

[45] A.R. Macrae. Lipase-catalyzed interesterification of oils and fats. *J. Am. Oil Chem. Soc.*, 60:291–294, 1983.

[46] J. Harwood. The versatility of lipases for industrial uses. *Trends Biotechnol.*, 14:125–126, 1989.

[47] Y.-Y. Linko, M. Lämsä, Y. Wu, E. Uosukainen, J. Seppälä, and P. Linko. Biodegradable products by lipase biocatalysis. *J. Biotechnol.*, 66:41–50, 1998.

[48] A.L. Margolin. Enzymes in synthesis of chiral drugs. *Enzyme Microb. Technol.*, 15:266–280, 1993.

[49] M.T. Reetz and K.-E. Jaeger. Enantioselective enzymes for organic synthesis created by directed evolution. *Chem. Eur. J.*, 6:407–412, 2000.

[50] L. Sarda and P. Desnuelle. Action de la lipase pancréatique sur les esters en émulsion. *Biochim. Biophys. Acta*, 30:513–521, 1958.

[51] A.M. Brzozowski, U. Derewenda, Z.S. Derewenda, G.G. Dodson, D.M. Lawson, J.P. Turkenburg, F. Bjorkling, B. Huge-Jensen, S.A. Patkar, and L. Thim. A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature*, 351:491–494, 1991.

[52] P. Grochulski, Y. Li, J.D. Schrag, F. Bouthillier, P. Smith, D. Harrison, B. Rubin, and M. Cygler. Insights into interfacial activation from an open structure of *Candida rugosa* lipase. *J. Biol. Chem.*, 268:12843–12847, 1993.

[53] J.D. Schrag and M. Cygler. 1.8 Å refined structure of the lipase from *Geotrichum candidum*. *J. Mol. Biol.*, 230:575–591, 1993.

[54] P. Grochulski, Y. Li, J.D. Schrag, and M. Cygler. Two conformational states of *Candida rugosa* lipase. *Protein Sci.*, 3:82–91, 1994.

[55] U. Derewenda, L. Swenson, R. Green, Y. Wei, S. Yamaguchi, R. Joerger, M.J. Haas, and Z.S. Derewenda. Current progress in crystallographic studies of new lipases from filamentous fungi. *Protein Eng.*, 7:551–557, 1994.

[56] M.E.M. Noble, A. Cleasby, L.N. Johnson, M.R. Egmond, and L.G.J. Frenken. The crystal structure of triacylglycerol lipase from *Pseudomonas glumae* reveals a partially redundant catalytic aspartate. *FEBS Lett.*, 331:123–128, 1993.

[57] D. Lang, B. Hofmann, L. Haalck, H.-J. Hecht, F. Spener, R.D. Schmid, and D. Schomburg. Crystal structure of a bacterial lipase from *Chromobacterium viscosum* ATCC 6918 refined at 1.6 Å resolution. *J. Mol. Biol.*, 259:704–717, 1996.

[58] J.D. Schrag, Y. Li, M. Cygler, D. Lang, T. Burgdorf, H.-J. Hecht, R. Schmid, D. Schomburg, T.J. Rydel, and J.D. Oliver *et al.* The open conformation of a *Pseudomonas* lipase. *Structure*, 5:187–202, 1997.

[59] M. Nardini, D.A. Lang, K. Liebeton, K.-E. Jaeger, and B.W. Dijkstra. Crystal structure of *Pseudomonas aeruginosa* lipase in the open conformation. *J. Biol. Chem.*, 275:31219–31225, 2000.

[60] J. Hermoso, D. Pignol, S. Penel, M. Roth, C. Chapus, and J.C. Fontecilla-Camps. Neutron crystallographic evidence of lipase-colipase complex activation by a micelle. *EMBO J.*, 16:5531–5536, 1997.

[61] D. Pignol, L. Ayvazian, B. Kerfelec, P. Timmins, I. Crenon, J. Hermoso, J.C. Fontecilla-Camps, and C. Chapus. Critical role of micelles in pancreatic lipase activation revealed by small angle neutron scattering. *J. Biol. Chem.*, 275:4220–4224, 2000.

[62] O.G. Berg, Y. Cajal, G.L. Butterfoss, R.L. Grey, M.A. Alsina, B.-Z. Yu, and M.K. Jain. Interfacial activation of triglyceride lipase from *Thermomyces (Humicola) lanuginosa*: kinetic parameters and a basis for control of the lid. *Biochemistry*, 37:6615–6627, 1998.

[63] Y. Cajal, A. Svendsen, V. Girona, S.A. Patkar, and M.A. Alsina. Interfacial control of lid opening in *Thermomyces lanuginosa* lipase. *Biochemistry*, 39:413–423, 2000.

[64] L.B. Lester and J.D. Scott. Anchoring and scaffolding proteins for kinases and phosphatases. *Recent Prog. Horm. Res.*, 52:409–430, 1997.

[65] M. Colledge and J.D. Scott. AKAPs: from structure to function. *Trends Cell Biol.*, 9:216–221, 1999.

[66] J.D. Scott, M.L. Dell'Acqua, I.D.C. Fraser, S.J. Tavalin, and L.B. Lester. Coordination of cAMP signaling events through PKA anchoring. *Adv. Pharmacol.*, 47:175–207, 2000.

[67] K. Dodge and J.D. Scott. AKAP79 and the evolution of the AKAP model. *FEBS Lett.*, 476:58–61, 2000.

[68] M.G. Newlon, M. Roy, D. Morikis, Z.E. Hausken, V. Coghlan, J.D. Scott, and P.A. Jennings. The molecular basis for protein kinase A anchoring revealed by solution NMR. *Nat. Struct. Biol.*, 6:222–227, 1999.

[69] P. Banky, M.G. Newlon, M. Roy, S. Garrod, S.S. Taylor, and P.A. Jennings. Isoform-specific differences between the type Iα and IIα cyclic AMP-dependent protein kinase anchoring domains revealed by solution NMR. *J. Biol. Chem.*, 275:35146–35152, 2000.

[70] K. Miki and E.M. Eddy. Single amino acids determine specificity of binding of protein kinase A regulatory subunits by protein kinase A anchoring proteins. *J. Biol. Chem.*, 274:29057–29062, 1999.

[71] F.W. Herberg, A. Maleszka, T. Eide, L. Vossebein, and K. Tasken. Analysis of A-kinase anchoring protein (AKAP) interaction with protein kinase A (PKA) regulatory subunits: PKA isoform specificity in AKAP binding. *J. Mol. Biol.*, 298:329–339, 2000.

[72] A.S. Edwards and J.D. Scott. A-kinase anchoring proteins: protein kinase A and beyond. *Curr. Opin. Cell Biol.*, 12:217–221, 2000.

[73] V.M. Coghlan, B.A. Perrino, M. Howard, L.K. Langeberg, J.B. Hicks, W.M. Gallatin, and J.D. Scott. Association of protein kinase A and protein phosphatase 2B with a common anchoring protein. *Science*, 267:108–111, 1995.

[74] T.M. Klauck, M.C. Faux, K. Labudda, L.K. Langeberg, S. Jaken, and J.D. Scott. Coordination of three signaling enzymes by AKAP79, a mammalian scaffold protein. *Science*, 271:1589–1592, 1996.

[75] I.D.C. Fraser, M. Cong, J. Kim, E.N. Rollins, Y. Daaka, R.J. Lefkowitz, and J.D. Scott. Assembly of an A kinase-anchoring protein-β_2-adrenergic receptor complex facilitates receptor phosphorylation and signaling. *Curr. Biol.*, 10:409–412, 2000.

[76] D. Perahia and L. Mouawad. Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Computers Chem.*, 19:241–246, 1995.

[77] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.

[78] S. Hayward, A. Kitao, and N. Gō. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins*, 23:177–186, 1995.

[79] D. Janežič, R.M. Venable, and B.R. Brooks. Harmonic analysis of large systems. III. Comparison with molecular dynamics. *J. Comp. Chem.*, 16:1554–1566, 1995.

[80] D.W. Miller and D.A. Agard. Enzyme specificity under dynamic control: a normal mode analysis of α-lytic protease. *J. Mol. Biol.*, 286:267–278, 1999.

[81] B. Brooks and M. Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA*, 82:4995–4999, 1985.

[82] O. Marques and Y.-H. Sanejouand. Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, 23:557–560, 1995.

[83] L. Mouawad and D. Perahia. Motions in hemoglobin studied by normal mode analysis and energy minimization: evidence for the existence of tertiary T-like, quaternary R-like intermediate structures. *J. Mol. Biol.*, 258:393–410, 1996.

[84] S. Hayward, A. Kitao, and N. Gō. Harmonic and anharmonic aspects in the dynamics of BPTI: a normal mode analysis and principal component analysis. *Protein Sci.*, 3:936–943, 1994.

[85] J.-F. Gibrat and N. Gō. Normal mode analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion. *Proteins*, 8:258–279, 1990.

[86] M.M. Tirion and D. ben-Avraham. Normal mode analysis of G-actin. *J. Mol. Biol.*, 230:186–195, 1993.

[87] A. Thomas, M.J. Field, L. Mouawad, and D. Perahia. Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. *J. Mol. Biol.*, 257:1070–1087, 1996.

[88] A. Thomas, M.J. Field, and D. Perahia. Analysis of the low-frequency normal modes of the R state of aspartate transcarbamylase and a comparison with the T state models. *J. Mol. Biol.*, 261:490–506, 1996.

[89] J. Ma and M. Karplus. Ligand-induced conformational changes in *ras* p21: a normal mode and energy minimization analysis. *J. Mol. Biol.*, 274:114–131, 1997.

[90] A. Thomas, K. Hinsen, M.J. Field, and D. Perahia. Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study. *Proteins*, 34:96–112, 1999.

[91] H. Kikuchi, H. Wako, K. Yura, M. Gō, and M. Mimuro. Significance of a two-domain structure in subunits of phycobiliproteins revealed by the normal mode analysis. *Biophys. J.*, 79:1587–1600, 2000.

[92] R. Diamond. On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Cryst. A*, 46:425–435, 1990.

[93] A. Kidera and N. Gō. Normal mode refinement: crystallographic refinement of protein dynamic structure. I. Theory and test by simulated diffraction data. *J. Mol. Biol.*, 225:457–475, 1992.

[94] B.R. Brooks, D. Janežič, and M. Karplus. Harmonic analysis of large systems. I. Methodology. *J. Comp. Chem.*, 16:1522–1542, 1995.

[95] I.K. M^{c}Donald and J.M. Thornton. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, 238:777–793, 1994.

[96] D. Bordo and P. Argos. The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. *J. Mol. Biol.*, 243:504–519, 1994.

[97] S.H. Bryant and C.E. Lawrence. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions. *Proteins*, 9:108–119, 1991.

[98] V.Z. Spassov and B.P. Atanasov. Spatial optimization of electrostatic interactions between the ionized groups in globular proteins. *Proteins*, 19:222–229, 1994.

[99] F. Avbelj and J. Moult. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34:755–764, 1995.

[100] P.H. Maccallum, R. Poet, and E.J. Milner-White. Coulombic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of α-helices and antiparallel β-sheet. A new method for analysing the forces between hydrogen bonding groups in proteins includes all the Coulombic interactions. *J. Mol. Biol.*, 248:361–373, 1995.

[101] P.H. Maccallum, R. Poet, and E.J. Milner-White. Coulombic attractions between partially charged main-chain atoms stabilise the right-handed twist found in most β-strands. *J. Mol. Biol.*, 248:374–384, 1995.

[102] J. Antosiewicz, S.T. Wlodek, and J.A. McCammon. Acetylcholinesterase: role of the enzyme's charge distribution in steering charged ligands toward the active site. *Biopolymers*, 39:85–94, 1996.

[103] T. Selzer and G. Schreiber. Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction. *J. Mol. Biol.*, 287:409–419, 1999.

[104] F.B. Sheinerman, R. Norel, and B. Honig. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 10:153–159, 2000.

[105] J. Warwicker and H.C. Watson. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.*, 157:671–679, 1982.

[106] M.K. Gilson, K.A. Sharp, and B.H. Honig. Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comp. Chem.*, 9:327–335, 1987.

[107] A.H. Elcock and J.A. McCammon. Evidence for electrostatic channeling in a fusion protein of malate dehydrogenase and citrate synthase. *Biochemistry*, 35:12652–12658, 1996.

[108] R.C. Tan, T.N. Truong, and J.A. McCammon. Acetylcholinesterase: electrostatic steering increases the rate of ligand binding. *Biochemistry*, 32:401–403, 1993.

[109] A.H. Elcock, R.R. Gabdoulline, R.C. Wade, and J.A. McCammon. Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin. *J. Mol. Biol.*, 291:149–162, 1999.

[110] R.R. Gabdoulline and R.C. Wade. On the protein-protein diffusional encounter complex. *J. Mol. Recognit.*, 12:226–234, 1999.

[111] J. Warwicker. Model for the differential stabilities of rhinovirus and poliovirus to mild acidic pH, based on electrostatics calculations. *J. Mol. Biol.*, 223:247–257, 1992.

[112] C.J. Gibas, S. Subramaniam, J.A. McCammon, B.C. Braden, and R.J. Poljak. pH dependence of antibody/lysozyme complexation. *Biochemistry*, 36:15599–15614, 1997.

[113] M. Schaefer, H.W.T. van Vlijmen, and M. Karplus. Electrostatic contributions to molecular free energies in solution. *Adv. Protein Chem.*, 51:1–57, 1998.

[114] J. Lamotte-Brasseur, V. Lounnas, X. Raquet, and R.C. Wade. pK_a calculations for class A β-lactamases: influence of substrate binding. *Protein Sci.*, 8:404–409, 1999.

[115] D. Livesay, S. Linthicum, and S. Subramaniam. pH dependence of antibody: hapten association. *Mol. Immunol.*, 36:397–410, 1999.

[116] E. Błachut-Okrasińska, B. Lesyng, J.M. Briggs, J.A. McCammon, and J.M. Antosiewicz. Poisson-Boltzmann model studies of molecular electrostatic properties of the cAMP-dependent protein kinase. *Eur. Biophys. J.*, 28:457–467, 1999.

[117] J. Trylska, J. Antosiewicz, M. Geller, C.N. Hodge, R.M. Klabe, M.S. Head, and M.K. Gilson. Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. *Protein Sci.*, 8:180–195, 1999.

[118] M. Mandel. The effect of temperature on the proton magnetic resonance spectra of ribonuclease, oxidized ribonuclease, and lysozyme. *Proc. Natl. Acad. Sci. USA*, 52:736–741, 1964.

[119] M. Mandel. Proton magnetic resonance spectra of some proteins. I. Ribonuclease, oxidized ribonuclease, lysozyme, and cytochrome $_c$. *J. Biol. Chem.*, 240:1586–1592, 1965.

[120] D.H. Meadows, J.L. Markley, J.S. Cohen, and O. Jardetzky. Nuclear magnetic resonance studies of the structure and binding sites of enzymes, I. histidine residues. *Proc. Natl. Acad. Sci. USA*, 58:1307–1313, 1967.

[121] R. Richarz and K. Wüthrich. High-field ^{13}C nuclear magnetic resonance studies at 90.5 MHz of the basic pancreatic trypsin inhibitor. *Biochemistry*, 17:2263–2269, 1978.

[122] D. Bashford and M. Karplus. pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*, 29:10219–10225, 1990.

[123] C.J. Gibas, P. Jambeck, and S. Subramaniam. Continuum electrostatic methods applied to pH-dependent properties of antibody-antigen association. *Methods*, 20:292–309, 2000.

[124] M.E. Davis, J.D. Madura, B.A. Luty, and J.A. McCammon. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian dynamics program. *Comput. Phys. Comm.*, 62:187–197, 1991.

[125] C. Tanford and J.G. Kirkwood. Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.*, 79:5333–5339, 1957.

[126] M.K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins*, 4:7–18, 1988.

[127] C. Tanford and R. Roxby. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry*, 11:2192–2198, 1972.

[128] M.K. Gilson. Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins. *Proteins*, 15:266–282, 1993.

[129] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, and D.L. Wheeler. GenBank. *Nucleic Acids Res.*, 28:15–18, 2000.

[130] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28:45–48, 2000.

[131] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[132] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–603, 1996.

[133] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[134] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, and M.D.R. Croning *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, 29:37–40, 2001.

[135] J. Schultz, R.R. Copley, T. Doerks, C.P. Ponting, and P. Bork. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, 28:231–234, 2000.

[136] K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. *Nucleic Acids Res.*, 27:215–219, 1999.

[137] A. Bateman, E. Birney, R. Durbin, S.E. Eddy, K.L. Howe, and E.L.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res.*, 28:263–266, 2000.

[138] T.K. Attwood, M.D.R. Croning, D.R. Flower, A.P. Lewis, J.E. Mabey, P. Scordis, J.N. Selley, and W. Wright. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, 28:225–227, 2000.

[139] J.G. Henikoff, E.A. Greene, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the Blocks database servers. *Nucleic Acids Res.*, 28:228–230, 2000.

[140] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[141] W.C. Barker, J.S. Garavelli, H. Huang, P.B. McGarvey, B.C. Orcutt, G.Y. Srinivasarao, C. Xiao, L.-S.L. Yeh, R.S. Ledley, and J.F. Janda *et al.* The protein information resource (PIR). *Nucleic Acids Res.*, 28:41–44, 2000.

[142] T.A. Tatusova and T.L. Madden. BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.*, 174:247–250, 1999.

[143] D.G. Higgins, A.J. Bleasby, and R. Fuchs. CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.*, 8:189–191, 1992.

[144] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choise. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[145] S.R. Eddy. Multiple alignment using hidden Markov models. In C. Rawlings, D. Clark, R. Altman, J.C. Hunter, and L.C. Rawlings, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 114–120. AAAI Press, Menlo Park, 1995.

[146] E. Beitz. TEXshade: shading and labeling of multiple sequence alignments using LATEX2ϵ. *Bioinformatics*, 16:135–139, 2000.

[147] K. Hofmann and M. Baron. unpublished.

[148] T.L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. In C. Rawlings, D. Clark, R. Altman, L.C. Hunter, and L.C. Rawlings, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 21–29. AAAI Press, Menlo Park, 1995.

[149] W.N. Grundy, T.L. Bailey, C.P. Elkan, and M.E. Baker. Meta-MEME: motif-based hidden Markov models of protein families. *Comput. Appl. Biosci.*, 13:397–406, 1997.

[150] S.R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[151] Z.S. Derewenda and U. Derewenda. Relationships among serine hydrolases: evidence for a common structural motif in triacylglyceride lipases and esterases. *Biochem. Cell Biol.*, 69:842–851, 1991.

[152] H. Frauenfelder, S.G. Sligar, and P.G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.

[153] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, second edition, 1999.

[154] I.H. Shrivastava and M.S.P. Sansom. Simulations of ion permeation through a potassium channel: molecular dynamics of KcsA in a phospholipid bilayer. *Biophys. J.*, 78:557–570, 2000.

[155] L. Stella, E.E. Di Iorio, M. Nicotra, and G. Ricci. Molecular dynamics simulations of human glutathione transferase P1-1: conformational fluctuations of the apo-structure. *Proteins*, 37:10–19, 1999.

[156] R. Ishima and D.A. Torchia. Protein dynamics from NMR. *Nat. Struct. Biol.*, 7:740–743, 2000.

[157] B.L. Stoddard. New results using Laue diffraction and time-resolved crystallography. *Curr. Opin. Struct. Biol.*, 8:612–618, 1998.

[158] W.L. Hubbell, D.S. Cafiso, and C. Altenbach. Identifying conformational changes with site-directed spin labeling. *Nat. Struct. Biol.*, 7:735–739, 2000.

[159] B.A. Wallace. Conformational changes by synchrotron radiation circular dichroism spectroscopy. *Nat. Struct. Biol.*, 7:708–709, 2000.

[160] S. Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nat. Struct. Biol.*, 7:724–729, 2000.

[161] A. Engel and D.J. Müller. Observing single biomolecules at work with the atomic force microscope. *Nat. Struct. Biol.*, 7:715–718, 2000.

[162] T.E. Fisher, P.E. Marszalek, and J.M. Fernandez. Stretching single molecules into novel conformations using the atomic force microscope. *Nat. Struct. Biol.*, 7:719–724, 2000.

[163] H. Saibil. Conformational changes studied by cryo-electron microscopy. *Nat. Struct. Biol.*, 7:711–714, 2000.

[164] B. van den Berg, M. Tessari, G.H. de Haas, H.M. Verheij, R. Boelens, and R. Kaptein. Solution structure of porcine pancreatic phospholipase A_2. *EMBO J.*, 14:4123–4131, 1995.

[165] C. Yuan, I.-J.L. Byeon, Y. Li, and M.-D. Tsai. Structural analysis of phospholipase A_2 from functional perspective. 1. Functionally relevant solution structure and roles of the hydrogen-bonding network. *Biochemistry*, 38:2909–2918, 1999.

[166] C. Dominguez, C. Sebban-Kreuzer, O. Bornet, B. Kerfelec, C. Chapus, and F. Guerlesquin. Interactions of bile salt micelles and colipase studied through intermolecular nOes. *FEBS Lett.*, 482:109–112, 2000.

[167] P.R. Selvin. The renaissance of fluorescence resonance energy transfer. *Nat. Struct. Biol.*, 7:730–734, 2000.

[168] Y. Suzuki, T. Yasunaga, R. Ohkura, T. Wakabayashi, and K. Sutoh. Swing of the lever arm of a myosin motor at the isomerization and phosphate-release steps. *Nature*, 396:380–383, 1998.

[169] A. Goldman. How to make my blood boil. *Structure*, 3:1277–1279, 1995.

[170] K.S.P. Yip, T.J. Stillman, K.L. Britton, P.J. Artymiuk, P.J. Baker, S.E. Sedelnikova, P.C. Engel, A. Pascuo, R. Chiaraluce, and V. Consalvi *et al.* The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure*, 3:1147–1158, 1995.

[171] G. Vogt, S. Woell, and P. Argos. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.*, 269:631–643, 1997.

[172] C. Vetriani, D.L. Maeder, N. Tolliday, K.S.-P. Yip, T.J. Stillman, K.L. Britton, D.W. Rice, H.H. Klump, and F.T. Robb. Protein thermostability above 100 ^{0}C: a key role for ionic interactions. *Proc. Natl. Acad. Sci. USA*, 95:12300–12305, 1998.

[173] J.H.G. Lebbink, S. Knapp, J. van der Oost, D. Rice, R. Ladenstein, and W.M. de Vos. Engineering activity and stability of *Thermotoga maritima* glutamate dehydrogenase. II: construction of a 16-residue ion-pair network at the subunit interface. *J. Mol. Biol.*, 289:357–369, 1999.

[174] L. Xiao and B. Honig. Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.*, 289:1435–1444, 1999.

[175] H.M. Verheij, A.J. Slotboom, and G.H. de Haas. Structure and function of phospholipase A_2. *Rev. Physiol. Biochem. Pharmacol.*, 91:91–203, 1981.

[176] B.W. Dijkstra, R. Renetseder, K.H. Kalk, W.G.J. Hol, and J. Drenth. Structure of porcine pancreatic phospholipase A_2 at 2.6 Å resolution and comparison with bovine phospholipase A_2. *J. Mol. Biol.*, 168:163–179, 1983.

[177] H.D. Beer, G. Wohlfahrt, J.E.G. McCarthy, D. Schomburg, and R.D. Schmid. Analysis of the catalytic mechanism of a fungal lipase using computer-aided design and structural mutants. *Protein Eng.*, 9:507–517, 1996.

[178] T. Tsukazaki, T.A. Chiang, A.F. Davison, L. Attisano, and J.L. Wrana. SARA, a FYVE domain protein that recruits Smad2 to the TGFβ receptor. *Cell*, 95:779–791, 1998.

[179] J. Massagué, S.W. Blain, and R.S. Lo. TGFβ signaling in growth control, cancer, and heritable disorders. *Cell*, 103:295–309, 2000.

[180] Y. Zhang and R. Derynck. Regulation of Smad signalling by protein associations and signalling crosstalk. *Trends Cell Biol.*, 9:274–279, 1999.

[181] J.L. Wrana and L. Attisano. The Smad pathway. *Cytokine Growth Factor Rev.*, 11:5–13, 2000.

[182] C.G. Burd and S.D. Emr. Phosphatidylinositol(3)-phosphate signaling mediated by specific binding to RING FYVE domains. *Mol. Cell*, 2:157–162, 1998.

[183] J.-M. Gaullier, A. Simonsen, A. D'Arrigo, B. Bremnes, H. Stenmark, and R. Aasland. FYVE fingers bind PtdIns(3)P. *Nature*, 394:432–433, 1998.

[184] V. Patki, D.C. Lawe, S. Corvera, J.V. Virbasius, and A. Chawla. A functional PtdIns(3)P-binding motif. *Nature*, 394:433–434, 1998.

[185] L. Wang, Y. Zhu, and K. Sharma. Transforming growth factor-β1 stimulates protein kinase A in mesangial cells. *J. Biol. Chem.*, 273:8522–8527, 1998.

[186] J. Joore, S. van de Water, M. Betist, A. van den Eijnden-van Raaij, and D. Zivkovic. Protein kinase A is involved in the induction of early mesodermal marker genes by activin. *Mech. Dev.*, 79:5–15, 1998.

[187] I.R. Gupta, T.D. Piscione, S. Grisaru, T. Phan, M. Macias-Silva, X. Zhou, C. Whiteside, J. L. Wrana, and N.D. Rosenblum. Protein kinase A is a negative regulator of renal branching morphogenesis and modulates inhibitory and stimulatory bone morphogenetic proteins. *J. Biol. Chem.*, 274:26305–26314, 1999.

[188] G.I. Patterson and R.W. Padgett. TGFβ-related pathways, roles in *Caenorhabditis elegans* development. *Trends Genet.*, 16:27–33, 2000.

[189] C.S. Zuker and R. Ranganathan. Signal transduction: the path to specificity. *Science*, 283:650–651, 1999.

[190] L.M. Luttrell, S.S.G. Ferguson, Y. Daaka, W.E. Miller, S. Maudsley, G.J. Della Rocca, F.-T. Lin, H. Kawakatsu, K. Owada, and D.K. Luttrell *et al.* β-arrestin-dependent formation of β_2 adrenergic receptor-Src protein kinase complexes. *Science*, 283:655–661, 1999.